Quantitative Methods in Neuroscience
(NEU466M)
Homework 2
Due: Tuesday Feb 6 by 5 pm in NHB 3.128 (or by 12:30pm in class if preferred)

In this assignment you will start working with Matlab, use index notation, and explore fitting, cross-validation, and simple statistical measures. General guidelines: Always read through each complete problem carefully before attempting any parts. Feel free to collaborate in groups of size 2-3, but always note the names of your collaborators on your submitted homework. For graphs, always clearly label your axes, use good color and symbol choices. Print out your matlab code, used to generate your results. For derivations you're asked to do with paper and pencil, feel free to turn in handwritten or typed-out work.

1) **Covariance and linear regression.** We are going to check whether Matlab is doing the right thing when generating linear fits, as well as exploring the difference between slope and covariance.

   a. Generate data `x = randn(N,1)` and `y = sqrt(2)*x + sqrt(2)*randn(N,1)` in Matlab (N=1000). Compute the sample covariance of this x, y data using the basic commands `sum, *, +, -` etc. only. Re-do the computation using `cov`. Compare (they should be the same).

   b. Apply the analytical solution for linear regression that we derived in class to make a linear fit to the data. Plot the data in a (x,y) scatterplot, and add the linear fit curve to this plot. Also compute Pearson's correlation coefficient for this data.

   c. This time generate a linear fit using `polyfit`, and add this fit to the previous plot (use a dashed line).

   d. Covariance, slope, and Pearson's correlation coefficient as a function of variance in the response variable: Double the variance of the independent noise component in the y-data (i.e., multiply the `randn` term by an additional factor of $\sqrt{2}$), while keeping the same model for x. What changes do you expect to see in the three quantities relative to your results from a., b., based on the analytical formulae from class? Explain and justify. What changes do you actually get?

   e. Covariance, slope, and Pearson's correlation coefficient as a function of variance in the explanatory variable: Do the same as d. but for x: Double the variance of the x-data (multiply the `randn` term by a factor of $\sqrt{2}$), while keeping the same model for y. What changes do you expect to see in the three quantities based on the analytical formulae from class? Explain and justify. What changes do you actually get?

2) **Nonlinear least-squares regression and cross-validation.** In class we noted that modeling data with a too-complex model can lead to "overfitting", and that using cross-validation can help avoid overfitting. In this problem will will explore this problem.

   a. Consider the model $y = 0.4 + 0.5\sin(2\pi x) + \xi$, where ξ is a Gaussian noise with unit variance and zero mean. Generate a vector x of length $N = 11$ in Matlab, with values evenly spaced over the interval $[0, 1]$. Generate two datasets `ytrain, yxval` using the command `ytrain = 0.4 +0.5*sin(2*pi*x)+randn(N,1)` and `yxval = 0.4 +0.5*sin(2*pi*x)+randn(N,1)`. Find the coefficients for a linear fit to this data using `polyfit`. Plot the data and the fit, clearly marking your axes, and choosing good symbol, line, and color options for all.

   b. Write a script file in Matlab to automatically generate and plot polynomial plots for fits to `ytrain` of degree 1 up to 11 (again with `polyfit`). Make a separate figure for each degree. Hint: you will need a for loop to run over the different degrees; for a given degree, you will need an inner for loop to sum up the terms in the polynomial, coefficient by coefficient.

   c. Within the script above, add a line to compute the squared error of the fit, for each degree. Save these squared errors in a vector, and generate a plot of squared error versus degree. This is the fit error on your training data.

   d. Finally, within the script above, also add a line to compute the squared error of the fit to the cross-validation data `yxval`, saving these squared errors in a different vector. This is the cross-validation error. Plot the squared cross-validation error versus degree on the same plot as in (c), but in a different color.

3) **Variance and covariance: theory.**

   This problem involves analytically manipulating terms (in symbolic form, with pencil and paper, using indices) related to means, variances, and correlations, so that you become comfortable working with these quantities.

   a. Recall that the sample mean or first moment of the variable x, over observations $\{x_1, \cdots, x_M\}$, is written $\langle x\rangle = \frac{1}{M}\sum_{i=1}^{M} x_i$. The second moment of x over these same observations is written $\langle x^2\rangle = \frac{1}{M}\sum_{i=1}^{M} x_i^2$. Show that the following two expressions for the variance of x are equivalent (i.e., show that equality holds):

$$\langle (x - \langle x\rangle)^2\rangle = \langle x^2\rangle - \langle x\rangle^2,$$

and show that the following two are equivalent expressions for the covariance between variables x and y:

$$\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle.$$

b. Show that minimizing the squared error of the linear fit:

$$\sum_{n=1}^{N} (x_n w^1 + w^0 - y_n)^2$$

with respect to the parameters w^0 and w^1 yields the following two equations:

$$w^0 + w^1 \langle x \rangle - y = 0 \tag{1}$$
$$w^0 \langle x \rangle + w^1 \langle x^2 \rangle - \langle xy \rangle = 0. \tag{2}$$

c. Derive the optimal solutions for w^0 and w^1 by simultaneously solving these two equations for the two unknown parameters. Hint: first multiply the first equation by $\langle x \rangle$.